Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis on the Tehran stock exchange

Sabaresan V[1], Bhuvaneshwaran S.K[2], Kavivendhan Ks[3]

[1]*Assistant Professor, Department of Information Technology, Agni College of Technology, Chennai*

[2,3]*UG Student, Department of Information Technology, Agni College of Technology, Chennai*

***Abstract* - Nowadays, a hot challenge for supermarket chains is to offer personalized services to their customers. Stock prediction, supplying the stock user an item list for the next purchase according to her current needs, is one of these services. Current approaches are not capable of capturing at the same time the different factors influencing the stock user decision process: co-occurrence, sequential, periodicity and re-currency of the purchased items. To this aim, we define a pattern Temporal Annotated Recurring Sequence able to capture simultaneously and adaptively all these factors. We define the method to extract and develop a predictor for the next stock named that, on top, is able to understand the level of the stock user and recommend the set of most necessary items. By adopting the supermarket chains could crop tailored suggestions for each individual stock user which in turn could effectively speed up their stock prediction sessions. A deep experimentation shows that they are able to explain the stock user purchase behavior, and that TBP outperforms the state-of-the-art competitors.**

***Index Terms* - Stock market, Trends prediction, Classification, Machine learning, Deep learning.**

I.INTRODUCTION

Stock Market prediction and analysis is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. Stock Market is the important part of the economy of the country and plays a vital role in the growth of the industry and commerce of the country that eventually affects the economy of the country. Both investors and industry are involved in the stock market and want to know whether some stock will rise or fall over a certain period of time. The stock market is the primary source for any company to raise funds for business expansions. It is based on the concept of demand and supply. If the demand for a company's stock is higher, then the company share price increases and if the demand for company's stock is low then the company share price decreases. Another motivation for research in this field is that it possesses many theoretical and experimental challenges. The most important of these is the Efficient Market Hypothesis (EMH), the hypothesis says that in an efficient market, stock market prices fully reflect available information about the market and its constituents and thus any opportunity of earning excess profit ceases to exist. One example of a big exchange is the New York Stock Exchange.

The task of stock prediction has always been a challenging problem for statistics experts and finance. The main reason behind this prediction is buying stocks that are likely to increase in price and then selling stocks that are probably to fall. Generally, there are two ways for stock market prediction. Fundamental analysis is one of them and relies on a company’s technique and fundamental information like market position, expenses and annual growth rates. The second one is the technical analysis method, which concentrates on previous stock prices and values.

This analysis uses historical charts and patterns to predict future prices. Stock markets were normally predicted by financial experts in the past time. However, data scientists have started solving prediction problems with the progress of learning techniques. Also, computer scientists have begun using machine learning methods to improve the performance of prediction models and enhance the accuracy of predictions. Employing deep learning was the next phase in improving prediction models with better performance. Stock market prediction is full of challenges, and data scientists usually confront some problems when they try to develop a predictive model. With the evolution of computer science, various new disciplines came into existence which provided better prediction models. One such discipline of computer science is Machine Learning. Over the years, machine learning has played a vital role in predictions. Predictions like workload management in cloud [1] [41-42], heart disease prediction [37], house rent price prediction [38], stock market price prediction [18] etc. were now possible with various techniques of machine learning. It helped in building new and improvised prediction models, which gave better results with lesser complexity. In context with stock market prediction, many researchers have been able to devise models for stock market prediction which uses various techniques of machine learning such as SVM (Support Vector Machine), Linear Regression, Random Forest, K-Nearest Neighbour (KNN), ANN, deep learning, LSTM, MLP, Boosted Decision Tree, Evolutionary algorithms and many more hybrid techniques which would be further discussed in this paper. This paper also discusses the challenges that are faced or can be faced by researchers while devising prediction models.

II. STOCK MARKET PREDICTION TECHNIQUES

Many models of prediction have been proposed till date to forecast the stock prices and stock market trends. Some of the machine learning techniques have been discussed in this paper. Table 1 displays the brief summary of all the techniques proposed by various researchers. All the techniques have been classified into various subcategories like classification techniques, regression techniques, ensemble algorithms, evolutionary techniques, deep learning, hybrid models and some other additional techniques.

A. Classification Techniques

1) Support Vector Machine (SVM): One of machine learning algorithms that possesses the desired features such as the decision function, usage of kernel methods and also the sparsity of the solution is known as the Support Vector Machine (SVM) technique.

Random Forest Classifier

Random forest classifier is a type of ensemble classifier and also a supervised algorithm. It basically creates a set of decision trees, that yields some result. The basic approach of random class classifier is to take the decision aggregate of random subset decision trees and yield a final class or result based on the votes of the random subset of decision trees.

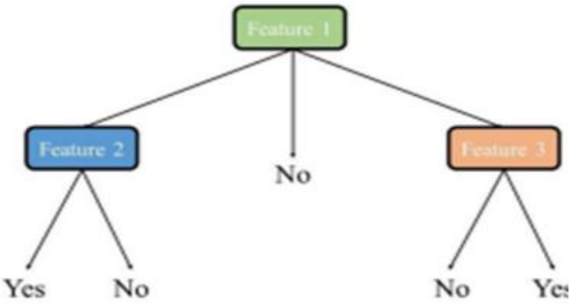
B. Random Forest Algorithm

Random forest algorithm is being used for the stock market prediction. Since it has been termed as one of the easiest to use and flexible machine learning algorithms, it gives good accuracy in the prediction. This is usually used in the classification tasks. Because of the high volatility in the stock market, the task of predicting is quite challenging. In stock market prediction we are using a random forest classifier which has the same hyperparameters as the decision tree. The decision tool has a model similar to that of a tree. It takes the decision based on possible consequences, which includes variables like event outcome, resource cost, and utility. The random forest algorithm represents an algorithm where it randomly selects different observations and features to build several decision trees and then takes the aggregate of the several decision trees outcomes. The data is split into partitions based on the questions on a label or an attribute. The data set we used was from the previous year’s stock markets collected from the public database available online, 80 % of data was used to train the machine and the rest 20 % to test the data. The basic approach of the supervised learning model is to learn the patterns and relationships in the data from the training set and then reproduce them for the test data.

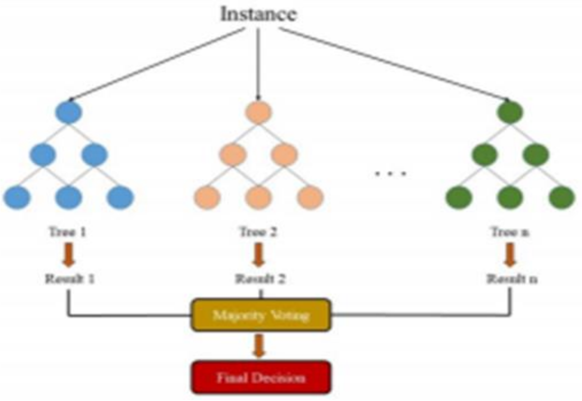
III. PREDICTION MODELS

In this study, we use nine machine learning methods (Decision Tree, Random Forest, Adaboost, XGBoost, SVC, Naïve Bayes, KNN, Logistic Regression and ANN) and two deep learning algorithms (RNN and LSTM).

A. Decision Tree: Decision Tree is a popular supervised learning approach employed for both regression and classification problems. The purpose is to make a model which is able to predict a target value by learning easy decision rules formed from the data features. There are some advantages of using this method like being easy to interpret and understand or Able to work out problems with multi-outputs; in contrast, creating over-complex trees that results in overfitting is a common disadvantage. A schematic illustration of Decision Tree is shown in Figure.



B. Random Forest: Great number of decision trees make a random forest model. The method simply averages the prediction result of trees, which is called a forest. Also, this model has three random concepts, randomly choosing training data when making trees, selecting some subsets of features when splitting nodes and considering only a subset of all features for splitting each node in each simple decision tree. During training data in a random forest, each tree learns from a random sample of the data points. A schematic illustration of Random Forest is indicated in Figure.



C. Adaboost: Boosting methods are a group of algorithms which convert weak learners to a powerful learner. The method is an ensemble for improving the model predictions of any learning algorithm. The concept of boosting is to sequentially train weak learners in order to modify their past prediction. AdaBoost is a meta-estimator which starts by fitting a model on the main dataset before fitting additional copies of the model on the similar dataset. During the process, samples' weights are adapted based on the current prediction error, so the subsequent model concentrates more on difficult items.

D. XGBoost: XGBoost is an ensemble tree-based method, and the model applies the principle of boosting for weak learners. XGBoost was introduced for better speed and performance in comparison with other tree-based models. In-built cross validation ability, regularization for avoiding overfitting, efficient handling of missing data, catch awareness, tree pruning, and parallelized tree building are common advantages of XGBoost method.

IV. EXPERIMENTAL RESULTS

A. Classification metrics

F1-Score, Accuracy and Receiver Operating Characteristics- Area Under the Curve (ROC-AUC) metrics are employed to evaluate the performance of our models. For Computing F1- score and Accuracy, Precision and Recall must be evaluated by Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). These values are indicated in Equations 7 and 8.

Precision = TP/TP + FP	(7)
Recall = TP/TP + FN	(8)

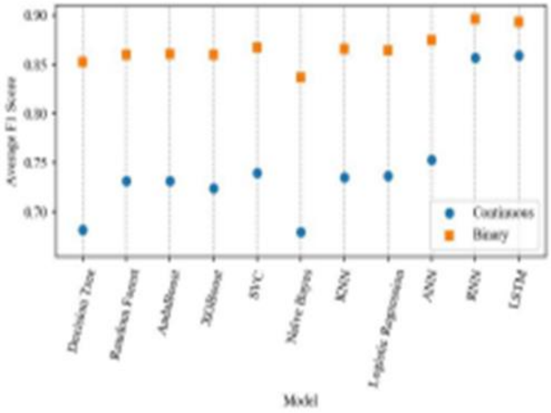
By calculation of above equations, F1-Score and Accuracy are defined in Equations 9 and 10.

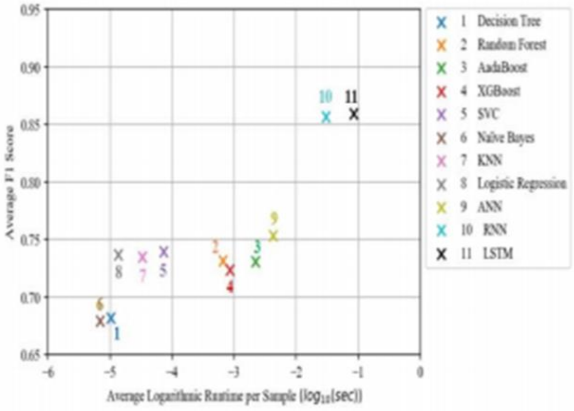
Accuracy = TP + TN/TP + FP + TN + FN	(9)
F 1 − Score = 2 Precision Recall Precision + (10) Re call	(10)

Among classification metrics, Accuracy is a good metric, but it is not enough for all classification problems. It is often necessary to look at some other metrics to make sure that the model is reliable. F1-Score might be a better metric to employ if results need to achieve a balance between Recall and Precision, especially when there is an uneven class distribution. ROC-AUC is another powerful metric for classification problems and is calculated based on the area under the ROC-AUC curve from prediction scores.

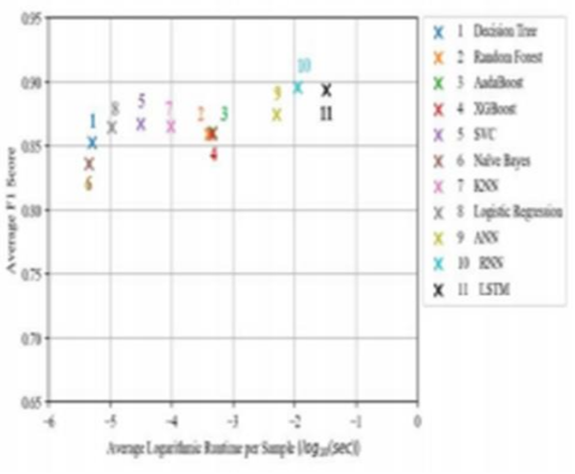
B. Results:

For training machine learning models, we implement the following steps: normalizing features (just for continuous data), randomly splitting the main dataset into train data and test data (30% of dataset was assigned to the test part), fitting the models and evaluating them by validation data (and "early stopping") to prevent overfitting, and using metrics for final evaluation with test data. Creating deep models is different from machine learning when the input values must be three dimensional (samples, time_steps, features); so, we use a function to reshape the input values. Also, weight regularization and dropout layers are employed to prevent overfitting here. All coding processes in this study are implemented by python with Scikit Learn and Kears library. Based on extensive experimental works by deeming the approaches, the following outcomes are obtained: In the first approach, continuous data for the features is used, and Tables 6-8 show the result of this method. For each model, the prediction performance is evaluated by the three metrics. Also, the best tuning parameter for all models (except Naïve Bayes and Logistic Regression) is reported. For achieving a better image of experimental works, Figure 14 is made to indicate the average of F1-score based on average running time through the stock market groups. It can be seen that Naive-Bayes and Decision Tree are least accurate (approximately 68%) while RNN and LSTM are top predictors (roughly 86%) with a considerable difference compared to other models. Indeed, the running time of those superiors is more than other algorithms. In the second approach, binary data for the features is employed, and Tables 9-11 demonstrate the result of this way. The structure and experimental works here are similar to the first approach except inputs where we use an extra layer to convert continuous data to binary one based on the nature and property of the features. Similarly, for better understanding, Figure 15 is made to show the average of F1- score based on average running time through the stock market groups. It is clear that there is a significant improvement in the prediction performance of all models in comparison with the first approach, and this achievement is obviously shown in Figure 16. There is no change in the inferior methods (Naive-Bayes and Decision Tree with roughly 85% F1-score) and the superior predictors (RNN and LSTM with approximately 90% F1-score), but the difference between them becomes less by binary data. Also, the prediction process for all models is faster in the second approach than others because of using large amounts of epochs and values related to some days before.





As a prominent result, deep learning methods (RNN and LSTM) show a powerful ability to predict stock movement in both approaches, especially for continuous data when the performance of machine learning models is so weaker than binary methods. However, the running time of those is always Overall, it is obvious that all the prediction models perform well when they are trained with continuous values (up to 67%), but the models' performance is remarkably improved when they are trained with binary data (up to 83%). The result behind this improvement is interpreted as follows: an extra layer is employed in the second approach, and the duty of the layer is comparing each current continuous value (at time t) with previous value (at time t-1). So, the future up or down trend is identified and when binary data is given as the input values to the predictors, we enter data with a recognized trend based on each feature's property.



This critical layer is able to convert non-stationary values in the first approach to trend deterministic values in the second one, and algorithms must find the correlation between input trends and output movement as an easier prediction task. Despite noticeable efforts to find valuable studies on the same stock market, there is not any significant paper to report, and this deficiency is one of the novelty of this research. We believe that this paper can be a baseline to compare for future studies.

V.CONCLUSIONS

The purpose of this study was the prediction task of stock market movement by machine learning and deep learning algorithms. Four stock market groups, namely diversified financials, petroleum, non-metallic minerals and basic metals, from Tehran stock exchange were chosen, and the dataset was based on ten years of historical records with ten technical features. Also, nine machine learning models (Decision Tree, Random Forest, Adaboost, XGBoost, SVC, Naïve Bayes, KNN, Logistic Regression and ANN) and two deep learning methods (RNN and LSTM) were employed as predictors. We supposed two approaches for input values to models, continuous data and binary data, and we employed three classification metrics for evaluations. Our experimental works showed that there was a significant improvement in the performance of models when they used binary data instead of continuous one. Indeed, deep learning algorithms (RNN and LSTM) were our superior models in both approaches.

REFERENCE

[1] Murphy, John J. Technical analysis of the financial markets: A comprehensive guide to trading methods and applications. Penguin, 1999

[2] Turner, Toni. A Beginner's Guide to Day Trading Online 2nd Edition. Simon and Schuster, 2007.

[3] Chen, Yingjun, and Yongtao Hao. "A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction." Expert Systems with Applications 80 (2017): 340-355.

[4] Basak, Suryoday, et al. "Predicting the direction of stock market prices using tree-based classifiers." The North American Journal of Economics and Finance 47 (2019): 552-567

[5] Araújo, Ricardo De A., and Tiago AE Ferreira. "A morphological-rank-linear evolutionary method for stock market prediction." Information Sciences 237 (2013): 3-17.

[6] Weng, Bin, et al. "Macroeconomic indicators alone can predict the monthly closing price of major US indices: Insights from artificial intelligence, time-series analysis and hybrid models." Applied Soft Computing 71 (2018):685-697.

[7] Baek, Yujin, and Ha Young Kim. "ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module." Expert Systems with Applications 113 (2018): 457-480.

[8] Patel, Jigar, et al. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques." Expert systems with applications 42.1 (2015): 259-268

[9] Majhi, Ritanjali, et al. "Efficient prediction of stock market indices using adaptive bacterial foraging optimization (ABFO) and BFO based techniques." Expert Systems with Applications 36.6 (2009): 10097-10104.

[10] Sun, Jie, and Hui Li. "Financial distress prediction using support vector machines: Ensemble vs. individual." Applied Soft Computing 12.8 (2012): 2254- 2265.

[11] Patel, Jigar, et al. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques." Expert systems with applications 42.1 (2015): 259-268